WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2009; **10**:912–931
Published online 15 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/wcm.802

Performance evaluation of utility-based scheduling schemes with QoS guarantees in IEEE 802.16/WiMAX systems

Razvan Pitic[1], Federico Serrelli[1], Simone Redana[2] and Antonio Capone[1*,†]

[1]*Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy*
[2]*Nokia Siemens Networks, Munich, Germany*

Summary

One of the most important benefits that WiMAX technology brings, the ability to provide differentiated quality of service (QoS) guarantees, could also prove to be the largest problem for system designers, because scheduling mechanisms able to cope with these demands have not been explicitly defined in the standard. In order to facilitate the understanding of how various scheduling schemes perform in a real system, we present here a detailed performance evaluation of some utility-based scheduling algorithms, covering aspects like fairness and QoS provisioning. Through a series of extensive simulations, we analyse the ability of the scheduling schemes considered to strike a balance between fairness among users, or more restrictively, user QoS requirement satisfaction, and system efficiency maximization. Further, we show how several simple algorithms could be used as building blocks, constructing a powerful mechanism that allows the system designer to obtain any desired system behaviour, or even to dynamically change from one profile to another, depending on specific network-related conditions. More specifically, by combining the benefits of proportional fair (PF) scheduling with the highly desirable system capacity maximization, and also taking into account a peak-to-average (PTA) channel quality metric, we are able to define a rule that outperforms traditional scheduling schemes, copes with various network conditions and provides graceful service degradation. Our results indicate that, by exploiting the intrinsic properties of orthogonal frequency division multiple access (OFDMA) as well as the mechanisms of the WiMAX system that are not regulated by the standard, one could increase the system efficiency, while fully respecting the QoS guarantees imposed. The use of algorithms that provide graceful performance degradation is highly advisable, in order to be able to employ a non-conservative call admission control (CAC) mechanism, which further improves the overall spectral efficiency by maintaining the system close to saturation at all times. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: WiMax; opportunistic scheduling; OFDMA; utility functions; QoS

1. Introduction

The last decade has brought increasing interest in high-speed Internet access on a large scale, mainly due to the diversification and wide spread of multimedia applications and services. The evolution of wireless technology meant that broadband wireless access (BWA) became an economically viable

*Correspondence to: Antonio Capone, Dipartimento di Elettronica e Informazione—Politecnico di Milano, 34/5, via Ponzio, 20133, Milan, Italy.
†E-mail: capone@elet.polimi.it

solution for providing last mile Internet access, thanks to its versatile and cost-effective deployment possibilities.

This trend lead to the creation of the IEEE 802.16 Working Group, with the goal of producing a family of standards for interoperable fixed BWA, supporting high bit-rate voice, data and video services, full quality of service (QoS), as well as advanced security features. In order to provide certified interoperability between system components developed by original equipment manufacturer (OEMs) and assure a rapid adoption of the IEEE 802.16 standard, the Worldwide Microwave Interoperability (WiMAX) Forum was created as a non-profit industry trade organization. The main result of this initiative is the development of a unique subset of baseline features, referred to as 'System Profiles', which are meant to provide directives specifically aimed at ensuring worldwide compliant practical implementations of devices based on the WiMAX technology.

Although the standard provides a robust, QoS oriented, MAC protocol, the details of scheduling and resource reservation management are not standardized, providing an important mechanism for vendors to differentiate their products. This flexibility can be exploited by system designers through the implementation of advanced scheduling schemes that meet client requests in terms of QoS, provide differentiated service to clients belonging to different classes-of-service, or simply maximize radio resource utilization.

Moreover, a novel approach to scheduling allows for several objectives to be targeted at the same time, through the use of *utility functions*, a concept imported from the economics field [1]. A utility function groups together several objectives, expressed in terms of costs and profits, with the use of system and operator-defined variables. The resulting objective function, denoted utility, is typically expressed as the weighted summation of these partial objectives and is used as a discriminator in order to provide prioritized medium access to users. In other words, different utility functions can be used in order to obtain different system behaviours, depending on the specific interests of the service provider [2].

Algorithms based on utility functions offer an integrated way of enforcing system operator defined policies, which could aim either at meeting specific client demands related to QoS parameters (i.e. minimum guaranteed data-rate, maximum average delay, etc), or at enforcing economically oriented targets (such as maximizing system spectral efficiency).

The main aim of this paper is to provide an extensive performance evaluation of utility-based packet scheduling techniques in IEEE 802.16/WiMAX systems, with an emphasis on multi class-of-service QoS provisioning. Further, we present a tutorial overview of the various mechanisms related to the resource allocation process which are imposed by the standard and, more importantly, we highlight the flexible design options which are not standardized. The specific behaviour of each of the algorithms considered is analysed, in terms of QoS provisioning as well as system efficiency, under the imposed complexity constraints. Moreover, we devise a set of guidelines which allows system designers to obtain any desired response from the implemented WiMAX system by using the algorithms provided as building blocks for constructing powerful, highly customizable resource allocation mechanisms.

The rest of this paper is organized as follows: in Section 2 we present an overview of a typical WiMAX system, focusing on the resource allocation and QoS management mechanisms. In Section 3 we review the state-of-the-art of utility-based scheduling schemes and in Section 4 we provide the system model used. The results of our extensive simulations are presented and discussed in Section 5, while the conclusions are drawn in Section 6.

2. IEEE 802.16/WiMAX System Overview

WiMAX (Worldwide Interoperability for Microwave Access) is a broadband technology intended to provide wireless connectivity for fixed, nomadic, portable and mobile users. Defined for wireless networking, WiMAX's capabilities in transmission range and spectral efficiency should make this technology able to efficiently cover the last mile in unwired zones, mitigating the digital divide problem, where present, as well as guarantee high-speed wireless access in metropolitan area networks (Wireless MANs), as an alternative to more common wired broadband systems like cable and digital subscriber line (DSL).

The standard for the equipment interoperability and certification has been defined and published during the last years by the IEEE as 802.16 in different versions, depending on system architecture and requirements. The 802.16-2004 [3] version of the standard specifies the air interface for a fixed to nomadic, point-to-multipoint deployment, while the 802.16e-2005 standard [4] (implemented by the industry as Mobile

DOI: 10.1002/wcm

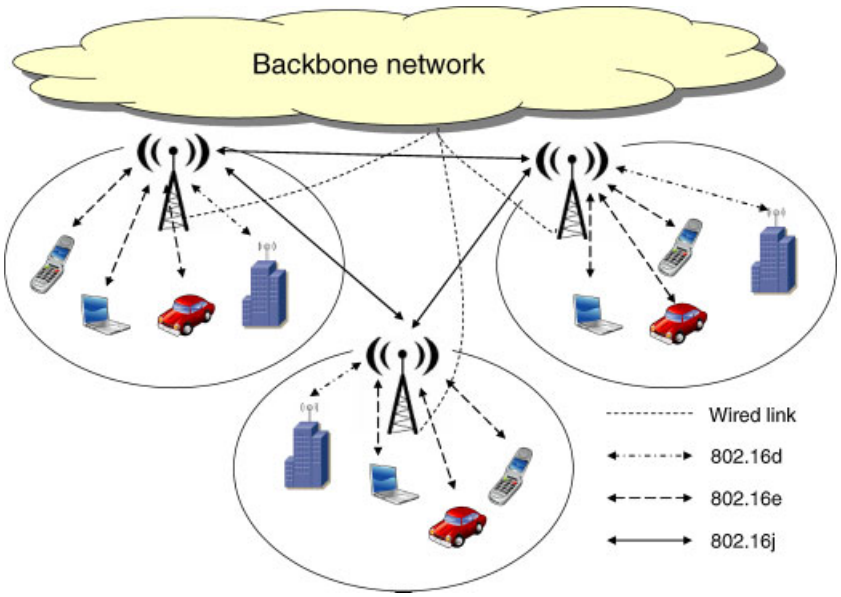


Fig. 1. A typical WiMAX system.

WiMAX) includes the previous versions and enables the support of higher mobility. Recently the standard has been extended to operate also in mesh mode (IEEE 802.16j), meaning that each node is able to forward data for other destinations and thus enabling the possibility for BSs (base stations) to interconnect through direct or relayed links in a mesh topology. A typical deployment scenario is shown in Figure 1.

Fixed and nomadic access is supported by using OFDM (orthogonal frequency division multiplexing) as modulation scheme in the 2–11 GHz frequency range, while the S-OFDMA (scalable orthogonal frequency division multiple access) guarantees connectivity for mobile users in 2–6 GHz band, offering a spectral efficiency of up to 2 bps/Hz for the downlink (DL) and 0.8 bps/Hz for the uplink (UL) [5].

In the next two sections we will detail some technical aspects of the standard that directly influence the resource allocation mechanism. For a more general overview of the WiMAX system, both fixed and mobile, good tutorials are provided in References [6] and respectively [7].

2.1. Resource Allocation

While for fixed access WiMAX uses OFDM in TDMA mode, allocating all subcarriers in a symbol to one specific flow or connection, in OFDMA the access of users to the medium can be diversified both in time and frequency, thus adding a degree of freedom in the resource scheduling process, allowing dynamic adaptation to variable channel conditions and introducing support for mobility.

In S-OFDMA, on which this work focuses, frequency resources are partitioned as follows: a group of contiguous logical subcarriers is termed as *subchannel* and represents the minimum granularity of resource allocation in the frequency domain. In the time dimension, consecutive OFDM symbols are grouped in frames. Knowing these definitions, the minimum resource allocation unit can be defined as the slot composed by one subchannel (in frequency) and a number of OFDM symbols[‡] (in time). At the beginning of each frame, the BS allocates all the OFDMA slots composing one frame among active users.

Scalability implies that the fast fourier transform (FFT) size is adapted depending on the available bandwidth, thus making variable the number of subcarriers in the OFDM symbol; however, also the number of subchannels is scaled consequently so that the amount of subcarriers composing one logical subchannel remains the same, maintaining this way unchanged all frequency-dependent channel characteristics when different bandwidth configurations are applied. This scalable subchannelization method is shown in Table I.

[‡]Depending on the subcarrier permutation rule applied.

DOI: 10.1002/wcm

Table I. WiMAX OFDM and S-OFDMA subchannelization.

Mobility	Standard	Access	Bandwidth [MHz]	Subcarriers [NFFT/Used/Data]	S-OFDMA: no. of SChs	S-OFDMA: no. of data SCs/SCh
Fixed to nomadic	802.16-2004	OFDM/TDMA	1.25-20	256/200/192	—	—
Portable to mobile	802.16e	S-OFDMA/TDMA	1.25	128/85/72	3	24
			5	512/420/360	15	
			10	1024/840/720	30	
			20	2048/1680/1440	60	

An important factor influencing the performance of S-OFDMA is the *subcarrier permutation rule*. In order to further combat the frequency selectivity of the wireless channel, the standard introduces distributed pseudo-random permutations (PUSC, FUSC) [4,8] in the mapping between logical and physical subcarriers. Mobility can cause fast variations of the wireless link quality and both the frequency selective scheduling and the dynamic adaptation of the modulation and coding scheme (MCS) become difficult due to the unreliability of users' channel quality information (CQI). Through subcarrier permutation it is possible to increase the frequency diversity within each single subchannel by reducing the correlation among the subcarriers composing it, thus enhancing the probability of correct decoding. This subcarrier scrambling has two other main advantages: the first is that the CQI could be considered substantially valid on the whole bandwidth and, if mobility is not excessive, this assumption can be extended over the entire frame. The second benefit is the possibility to randomize the interference coming from adjacent cells by setting differently the pseudo-random permutation applied in each site [9]. The downside is that by using this technique, the average quality of all subchannels becomes similar, which quasi nullifies the overall frequency diversity.

The standard also provides an adjacent subcarrier permutation. This scheme, called AMC (adaptive modulation and coding), maintains the same mapping between logical and physical subcarriers and gives the possibility of adapting the MCSs for each user depending on allocated subchannels and on the base of reliable and precise information about the frequency profile of the channel. It has been proved [10,11] that this type of opportunistic frequency selective scheduling can provide throughput maximization when low-mid mobility is considered due to the highly time-correlated channel.

The MAC frame (Figure 2) starts with a Preamble used for signal to interference (SINR) estimation and synchronization purposes. Control messages (FCH, DLMAP and ULMAP) are transmitted immediately after and carry all information about data allocation in the current frame in order to allow correct data

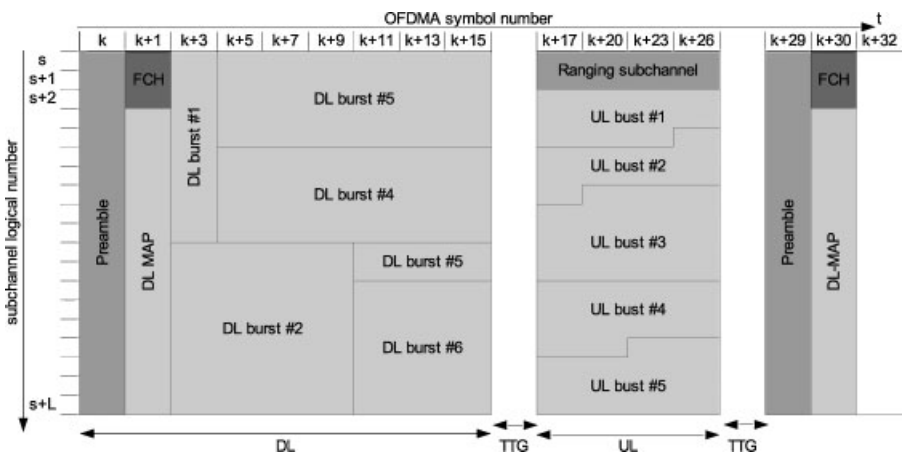


Fig. 2. Frame structure.

DOI: 10.1002/wcm

Table II. WiMAX scheduling services QoS parameters.

Scheduling service	Main QoS parameters
RT-CBR	Tolerated jitter, minimum reserved traffic rate, maximum latency
RT-VR	Maximum latency, minimum reserved traffic rate, maximum sustained traffic rate
DT-VR	Minimum reserved traffic rate, maximum sustained traffic rate
BE	—

localization and decoding. A typical frame duration is 5 ms and the split point between DL and UL is variable and could be changed on a frame by frame basis following some predefined DL/UL ratios. The DL data subframe can be divided in multiple permutation zones each using a different subcarrier permutation scheme. However, the standard stipulates that only a DL PUSC permutation zone is mandatory for carrying control messages and, optionally, data.

It should be noted that the DL PUSC permutation scheme imposes that data are mapped into rectangular data regions within a frame (see Figure 2); each region is formed by a number of contiguous subchannels (in frequency) and OFDM symbols (in time), which carry data pertaining to one or more users and employ the same MCS. This granularity should be taken into consideration when performing the frame mapping, so that the least possible amount of resources is wasted.

The necessity to transmit all control information at the beginning of each frame implies that all MAC procedures related to resource allocation must be accomplished by the BS within the previous frame interval. Such procedures include: link adaptation, frame mapping, power control and resource scheduling. Because of these tight time constraints, using a low-complexity scheduling scheme becomes a key factor in any real-system implementation, especially when complex differentiated QoS provisioning is required.

2.2. QoS Management

The WiMAX technology is intended to offer differentiated QoS provisioning. Each connection is associated a scheduling service which describes the QoS parameters to be provided by the resource allocation mechanism at the BS.

Each packet belonging to a data flow is mapped to a specific service flow by the convergence sublayer (CS). The CS resides at the top of the MAC layer and manages the mapping between MAC SDUs and transport connections. CS is responsible for packet classification and optional payload header suppression. When a new packet arrives at the CS, it is classified and associated to a specific unidirectional service flow, identified by a service flow ID (SFID), and mapped on the corresponding transport connection (CID). Each service flow is characterized by a set of QoS parameters such as maximum latency, minimum average throughput and tolerated jitter. The BS should guarantee that resource allocation is made respecting the imposed QoS requirements.

The standard supports four types of scheduling services: *Real-Time Constant Bit-Rate*, *Real-Time Variable Rate*, *Delay-Tolerant Variable Rate* and *Best Effort*. A set of QoS parameters is associated to each scheduling service; in Table II the typical QoS parameters for all scheduling services are listed.

3. Utility-based Scheduling

Due to the inherent unreliability of wireless channels, the packet scheduling process becomes the most important means by which end-to-end QoS could be guaranteed. Taking into account that a system operator also wants to maximize the use of the scarce radio resources, the quest for a scheduling scheme that provides full QoS support for various classes of service, while at the same time maximizing the system spectral efficiency, is taken to another level.

Historically, most of the scheduling schemes that were first proposed for WiMAX were adopted from single-carrier (SC) systems. While for the SC physical layer modes (i.e. WirelessMAN-SC and WirelessMAN-SCa) these algorithms are adequate, for the more advanced multi-carrier modes (i.e. WirelessMAN-OFDM and WirelessMAN-OFDMA) the results produced are suboptimal. This is mainly because OFDM/OFDMA systems have particularities which require totally different approaches in order to exploit their intrinsic properties.

The present digital modulation scheme of choice is orthogonal frequency division multiple access (OFDMA), considered the most advanced and

DOI: 10.1002/wcm

promising technology for the PHY layer. This is also reflected in the latest IEEE Standard 802.16e-2005 [4], which adopts it as its principal multi-access scheme. Seen as the natural evolution of OFDM for multiple access, OFDMA inherits the robustness to fast fading and interference of its predecessor, and adds several other benefits: reduced multi-user interference, flexible frequency reuse, subchannelization, multi-user diversity, better spectral efficiency etc. Even more importantly, the advent of OFDMA changes a fundamental concept of wireless system design: the fact that channel variation is considered a negative effect and is heavily combated using advanced techniques such as interleaving, power control, equalization, etc.

The traditional TDM scheduling algorithm, in single carrier systems, has been round robin (RR) [12]. It assigns time-slots to users in a sequential manner, independent of the channel conditions. A multi-service extension was developed, the weighted round robin (WRR) [13], which supports multiple service classes by sequentially assigning each user a number of time-slots dependent on the class it belongs to. Although these algorithms succeed in achieving a level of fairness among users, they provide very low spectral efficiency, a flaw which makes them unsuited for use in broadband networks.

Another class of algorithms, proportional fair (PF), is targeted at ensuring fair resource access for all users on a long time scale. The best known implementation of a PF algorithm is Qualcomm's high data rate (HDR) (CDMA2000/1xEV) standard. The principle behind this algorithm is to choose for transmission, at each time-slot t, the user i with the highest ratio between its respective instantaneous channel capacity R_i and average data rate $\overline{R}_i(t)$ over a sliding time window of size t_w:

$$i^* = \arg\max_i \frac{R_i}{\overline{R}_i} \tag{1}$$

where the average data rate can be updated from one time slot to the next using the following low-pass filter:

$$\overline{R}_i(t+1) = \left(1 - \frac{1}{t_w}\right)\overline{R}_i(t) + a_i(t)\frac{1}{t_w}R_i(t) \tag{2}$$

$a_i(t)$ is a boolean variable that is equal to 1 if user i was granted a transmission opportunity at time-slot t and 0 otherwise.

Different versions of PF algorithms have been adapted for multi-carrier systems, one of the best known being MPF [14]. These algorithms fail in a heterogeneous traffic environment due to their inability of providing QoS guarantees.

In a wireless system, users have statistically independent time-varying channels. This translates to different users experiencing peaks in their channel quality at different times. In a densely populated system, exploiting this effect by scheduling transmissions for users only when they have favourable channel conditions could lead to significant increase in the total system throughput. This effect has been called *multiuser diversity* [15] and forms the base of *opportunistic scheduling* (OS) [16]. In an OFDMA system this effect is more visible due to the fine granularity of the scheduling space (i.e. a large number of subcarriers and short symbol duration), which translates into higher variations of user channel quality, both in frequency and in time.

OS schemes use a cross-layer approach to MAC packet scheduling by leveraging the channel state information (CSI) retrieved from the physical layer in order to exploit multiuser diversity. A pure opportunistic approach always schedules for transmission the users experiencing the best channel quality, thus maximizing the system throughput. The main drawback of this approach is that it cannot provide any degree of fairness, which can lead to increased delays and even starvation for users with prolonged bad channel quality (i.e. located in deep fade areas).

To mitigate the problem of delay, several algorithms have been proposed. The best algorithms in this family take into account both channel conditions and QoS requirements in terms of stochastic packet delay bounds [17,18]. Such an algorithm is the Modified Largest Weighted Delay First (M-LWDF), which could be defined as follows: at each time instance, schedule for transmission the user satisfying the following condition:

$$i^* = \arg\max_i \rho_i W_i R_i \tag{3}$$

where ρ_i is a constant, W_i the head-of-line packet delay and R_i is the instantaneous channel capacity for user i. The proposed scheduler achieves throughput optimality, defined in Reference [18] as follows: a scheduling algorithm is throughput optimal if it is able to keep all queues stable if this is at all feasible to do with any scheduling algorithm.

Another algorithm from the same class is the exponential rule, proposed in Reference [19], which tries to equalize the weighted delays of all the queues when their differences are large. In a simplified form it

DOI: 10.1002/wcm

could be stated as follows:

$$i^* = \arg\max_i \; a_i \frac{R_i}{\overline{R}_i} exp\left(\frac{a_i W_i - \overline{aW}}{1 + \sqrt{\overline{aW}}}\right) \qquad (4)$$

where a_i is a weighting factor, R_i the instantaneous data rate supported by the channel for user i, $\overline{R}_i$ its respective mean data rate, W_i the head-of-line packet delay and $\overline{aW}$ is the arithmetic mean over all $a_i W_i$.

Using the exponential rule, when all queues have similar lengths, the user perceived channel conditions play a significant part. On the other hand, if one queue is much longer than others, then the queue length becomes dominant and the longer queue gets a higher chance to transmit. Hence, this algorithm balances the tradeoff between queue length and throughput. In other words, the exponential rule gracefully adapts from a proportionally fair one to one which balances the delays. The main downside of these scheduling algorithms is that their performance depends heavily on parameter settings, which is not desirable in a highly dynamic system where operator policies need to be changed depending on the traffic conditions of the network.

The IEEE 802.16 standard stipulates that the base station (BS) centrally allocates the available channels in each time slot to the various active subscriber stations (SSs) for both UL and DL, which in turn allocate these resources to the connections they manage at that time. In order to efficiently satisfy users requirements and at the same time achieve high spectral efficiency, a plethora of factors need to be considered for each user when making the scheduling decision: the instantaneous and average channel quality, packet backlog size, QoS parameters, average data-rate, etc.

Several scheduling algorithms have been tailored for IEEE 802.16 systems. An overview of the principles of joint scheduling and resource allocation in channel aware WiMAX systems (AMC mode) can be found in Reference [20]. Several classic algorithms from literature (such as Round Robin, PF, Max CINR) have been adapted to a Mobile WiMAX system and a brief comparison of their performance in terms of heterogeneous traffic support is presented in Reference [21]. A joint bandwidth allocation and connection admission control framework, which considers an optimization formulation based on a queuing analytical model, as well as a more practical iterative approach based on the the water-filling method, has been developed in Reference [22].

In a real life system, different subchannelization techniques (grouping several subcarriers into one subchannel) can be used in order to reach a trade off between the granularity of resource allocation and the overhead incurred by the channel quality measurement and feedback mechanism.

4. System Model

In this paper the system is modelled according to the specifications of Mobile WiMAX (based on the IEEE 802.16e standard). System parameters are summarized in Table III.

All scheduling algorithms have been evaluated and compared through system level simulations [23]. A fully standard compliant system simulator, based on ns-2 [24], implements all main functions of the MAC layer and manages multiple DL connections between the target BS and traced SSs. Statistics are extracted either from BS or SSs in order to assess the overall system performances.

Only the mandatory DL PUSC permutation mode is considered throughout this paper. The topology consists of a single cell and an intra-site frequency reuse factor of 3 is assumed, so that in each hexagonal sector the BS can employ one third of the 15 MHz available bandwidth. For 5 MHz the scaled FFT size is 512 and the resulting number of subchannels in each sector is 15.

Table III. System parameters.

Parameter	Value
Topology type	hexagonal, tri-sectorial
Centre frequency	2.3 GHz
Total available bandwidth in site	15 MHz
Frequency reuse	1:3
Available bandwidth in sector	5 MHz (frequency reuse 3 intra-site)
FFT size	512
Number of subchannels	15
Frame duration	5 ms
OFDM symbols per frame	47
Subcarrier permutation	PUSC
DL/UL ratio	(35:12) 1 OFDM symbol for frame preamble 6 OFDM symbols for FCH and MAPs 28 OFDM symbols for DL data 12 OFDM symbols for UL

DOI: 10.1002/wcm

Table IV. Physical layer model parameters.

Parameter	Value
BS Tx power/sector	35 dBm
BS antenna height	30 m
BS antenna pattern	65 deg (−3 dB) with 50 dB front-to-back ratio
BS antenna gain	17.5 dBi
SS antenna height	1.5 m
SS antenna pattern	Omnidirectional
SS noise figure	7 dB
SS mobility	3 km/h

The OFDMA slot size in DL PUSC configuration is 1 subchannel per 2 OFDM symbols so that the DL subframe is a 14×15 matrix and 210 slots are available for DL data transmissions in each frame.

The wireless channel is simulated through the use of pathloss, shadowing and fast fading components. Okumura–Hata model for urban environment is applied for median pathloss calculation and an additive lognormal shadowing is applied for each wireless link with null average and 8 dB standard deviation. These values are assumed constant for each link due to the limited duration of the simulation run time with respect to the typical slow fading dynamics. The fast fading is simulated for each wireless link by importing a channel trace with OFDM symbol resolution. The trace was generated offline using a link level simulator which implements an extension of the 3GPP Spatial Channel Model (SCME), developed in the context of the European WINNER project [25,26]. Physical layer and wireless channel modelling parameters are shown in Tables IV and V, respectively.

In order to evaluate the system performance, a fixed number of SSs have been positioned in the hexagonal sector so that in each simulation run the user average channel conditions are equally distributed, ranging from excellent to poor. For this purpose, the hexagonal sector under consideration has been divided in three

Table V. Propagation model parameters.

Parameter	Value
Pathloss model	Okumura–Hata
Pathloss slope	≈ 38 dB/decade
Shadowing model	Lognormal
Shadowing standard dev	8 dB
Fast fading	3GPP SCM

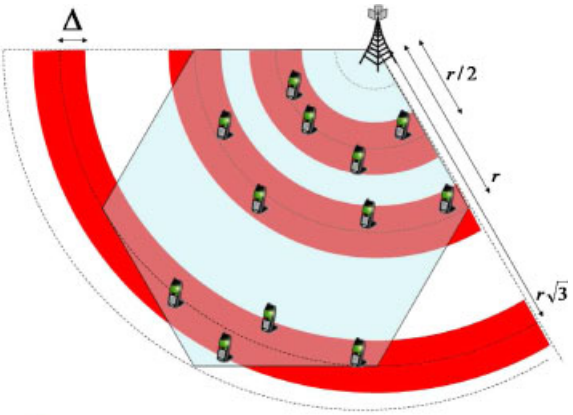


Fig. 3. Cell zones.

zones, defined as a function of the average distance from the BS (D_1, D_2 and D_3) and the zone width Δ, as shown in Figure 3.

An equal number of fully backlogged users have been randomly positioned in each zone. The network model parameters are listed in Table VI.

This assumption has been made in order to limit the number of simulation runs, but it is important to notice that due to shadowing and fast fading the actual channel quality of users can differ significantly even though they belong to the same zone. This implies that the duration and the overall number of simulation runs have to be large enough in order to guarantee the averaging of the involved statistical propagative components and the reliability of the measured statistics.

A total number of users greater than 20 should guarantee an adequate level of user diversity, so seven users per zone are simulated. A topology example for a single simulation run is depicted in Figure 4.

In Figure 5 the entire resource allocation process is represented. At the beginning of each frame the triggered scheduler allocates resources using the utility approach and builds a list of requests; required

Table VI. Network model parameters.

Parameter	Value
Cell radius r	500 m
Zone width Δ	$r/8 = 62.5$ m
Zone 1 average distance	$D_1 = r/2 = 250$ m
Zone 2 average distance	$D_2 = r = 500$ m
Zone 3 average distance	$D_3 = r\sqrt{3} = 866$ m
Number of SSs per zone	7
Traffic model	Full buffer

DOI: 10.1002/wcm

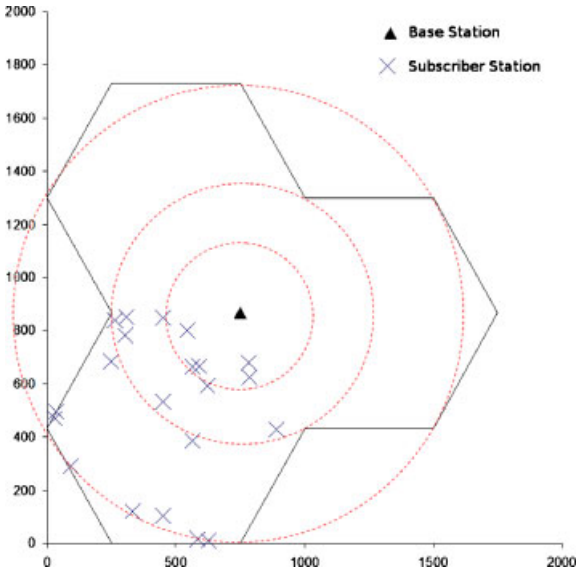


Fig. 4. Simulated network topology.

information, i.e. users' MCSs and the amount of data in each queue, are retrieved by the scheduler from the queues' manager and the link adaptation module.

Requests are passed to the frame mapper for creating rectangular data regions in the frame and obtaining the final frame structure. It is important to remark that due to the rectangular shape constraint, the percentage of the frame actually mapped depends on the particular request list and can vary in each frame [27]. Therefore a mapping algorithm able to map on average up to 90% of the frame has been implemented and a feedback on actually allocated slots is provided by the frame mapper when the frame structure is completed in order to update users' average throughput and utility functions.

Finally, just before transmission, packets are drawn from the queues and MAC PDUs are formed adding MAC overheads and applying fragmentation and packing if needed.

The Link Adaptation module is responsible for the adaptation of MCSs according to the CQI measured and reported by each SS. In this case the CQI corresponds to the SINR of the preamble field calculated through the so-called EESM (exponential effective SINR mapping) link to system interface [28]. The EESM formula is

$$\mathrm{SINR}_{\mathrm{PREAMBLE}}(t) = -\beta \ln\left(\frac{1}{N}\sum_{n=1}^{N}\exp\left(-\frac{\mathrm{SINR}_n(t)}{\beta}\right)\right) \tag{5}$$

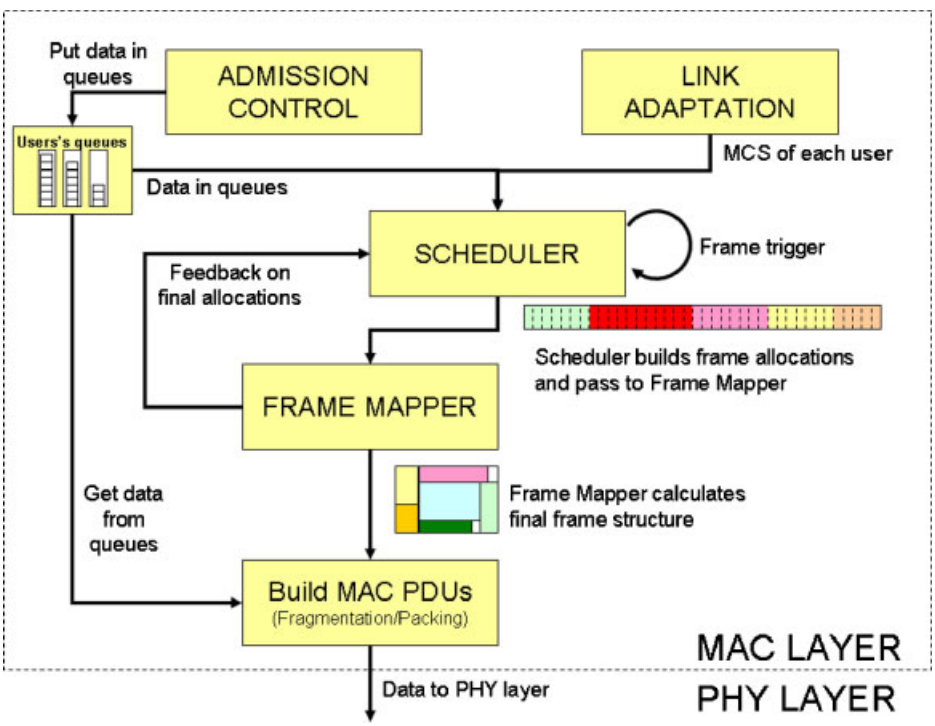


Fig. 5. MAC functional blocks.

DOI: 10.1002/wcm

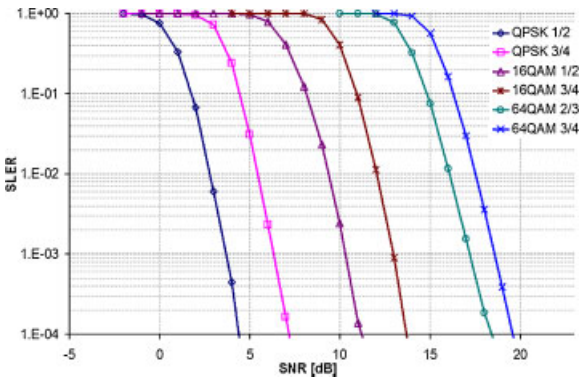


Fig. 6. Link quality curves.

Table VII. Modulation and coding schemes employed.

MCS	Required SNR [dB]
QPSK 1/2	2.9
QPSK 3/4	5.7
16QAM 1/2	9.6
16QAM 3/4	12.1
64QAM 2/3	16.1
64QAM 3/4	17.7

where N is the number of modulated subcarriers for preamble (120 for FFT size 512), $\mathrm{SINR}_n(t)$ is the SINR on the nth subcarrier at time t and β is a scaling factor tuned through link layer simulations for each possible MCS in order to allow the use of AWGN curves for link performance evaluation.

In order to select the most efficient modulation and coding scheme for transmission that would still assure a SLot Error Rate (SLER) lower than 10^{-2} for a given CQI value, the minimum required SNR thresholds have been extracted from the link performance curves [10] depicted in Figure 6 and are presented in Table VII.

5. Simulation Results

5.1. Performance Metrics

In this section different scheduling approaches are introduced and evaluated through system level simulations. The main scope is to verify and compare their capabilities in finding a trade-off between QoS provisioning and system throughput maximization. In particular we are interested in investigating the advantages coming from a careful design of utility functions in order to explicitly consider differentiated user QoS requirements in the scheduling metric calculation. Moreover, we want to assess the performance improvements achievable through the exploitation of *multiuser diversity*, leveraged in order to achieve an enhanced system utilization combined with QoS guarantees satisfaction.

Although in DL PUSC configuration the frequency diversity is deliberately nullified by the pseudo-random subcarrier permutation, the advantage coming from the exploitation of time diversity over an entire time window for QoS provisioning is large. For that reason all analysed utility functions are based on *minimum reserved traffic rate* (mRTR) guarantees because of the long-term nature of this QoS requirement. The mRTR parameter is expressed in bits per second and specifies the minimum amount of data to be transported on behalf of the service flow when averaged over time.

The following metrics are defined for performance evaluation:

- *Average System Throughput*: overall sector throughput averaged on all simulation runs.
- *Average User Throughput*: defined for each simulated user as the average throughput over the whole simulated session.
- *Long-Term User Satisfaction* (LT-US): defined for each simulated user as the ratio between the *Average User Throughput* and the mRTR (minimum reserved traffic rate). A value higher than 1 indicates that the user received sufficient resources considering the whole session. Although through this metric it is possible to verify the system behaviour with respect to long-term fairness and QoS provisioning, it is not enough in order to show the variability of the performance level perceived by users during the simulation in dependence, for instance, on channel quality variation.
- *Minimum Short-Term User Satisfaction* (mST-US): after dividing the user's session in time windows of duration TW (of 0.5 s), for each time window the ratio between the average throughput and the mRTR is calculated and the mST-US is defined as the minimum between these values over the whole session. In other words mST-US represents the minimum short-term QoS level perceived by the user during the session. This metric allows the evaluation of the scheduling algorithms' capabilities to provide QoS guarantees uniformly during the whole user's session and shows existing relations between time

DOI: 10.1002/wcm

varying quantities (e.g. channel condition) and perceived performance levels.

In the following, mST-US and LT-US will be used to define the system *satisfaction level*. If U is the set of all simulated users, for each possible value x of user satisfaction (US), either long or short-term, the *satisfaction level* represents the percentage of users exceeding that US level and can be defined as follows:

$$\text{Satisfaction\_level}(x) = \frac{\text{card}\{u \in U \mid \text{US}_u \geq x\}}{\text{card}\{U\}} \quad (6)$$

where card$\{U\}$ represents the cardinality of the set U.

In practice the Satisfaction_level(x) is the ratio between the number of users having a User Satisfaction degree greater than or equal to x and the total number of users.

5.2. Efficiency Maximization *Versus* Fairness

In wireless networks efficiency maximization is definitely one of the main objectives due to the scarcity of radio resources. Moreover, the unpredictable nature of the wireless channel makes essential the design and implementation of mechanisms able to react against a temporary lack of radio resources and exploit all the degrees of freedom provided by the system as efficiently as possible. The design of advanced scheduling schemes is one of the most effective ways to pursue this aim.

In the absence of any constraints or objectives related to the single user performance, the simplest way to maximize the efficiency is by using the maximum throughput (MT) scheduling algorithm. MT allocates every OFDMA slot to the user experiencing the best channel quality, thus assuring that all radio resources are employed using the most efficient transmission mode. In OFDM-based systems the AMC mechanism allows the use of different transmission modes, with efficiency levels proportional to the wireless channel quality (i.e. user's SINR).

In our case, due to the large user diversity, at each time instant there is at least one user experiencing a peak in channel quality and thus all slots are modulated with the most efficient MCS (64QAM 3/4). The resulting system throughput is

$$\Theta_{\text{MT}} = \frac{N_{\text{slot}} \cdot \text{bps}_{\text{64QAM3/4}}}{T_{\text{frame}}} \cdot \text{FO}_{\text{MT}} = \frac{210 \cdot 216}{0.005} \cdot 0.86 = 7.8\text{Mbps} \quad (7)$$

where N_{slot} is the number of OFDMA slots in a frame, $\text{bps}_{\text{64QAM3/4}}$ is the number of bits per slot that can be transmitted using the 64QAM3/4 MCS and T_{frame} represents the frame duration. The measured average frame occupation FO_{MT} is lower than 1 because the frame mapping algorithm has to achieve a trade-off between low complexity and the exhaustive quest for the optimal solution. In fact, Θ_{MT} represents the maximum achievable system throughput, with a corresponding spectral efficiency of 1.4 bps/Hz.

In a real WiMAX system the frame mapping problem, which can be considered a particular two-dimensional case of a more general NP-hard problem of operational research known in literature as the *Knapsack Problem* [29,30], can represent an important bottleneck and requires the implementation of greedy algorithms able to find a suboptimal solution in a finite time interval. Therefore, depending on allocation requests, it is not always possible to map the entire frame and a variable portion of resources remains often unallocated or some requests unfulfiled.

However, MT scheduling does not aim at establishing any sort of policy or priority among users on the base of the performance level actually perceived by each of them. Forcing a policy in the radio resource sharing could translate into different concepts (like fairness, minimum QoS guarantees, etc.) but, in any case, this will determine a trade-off between the spectral efficiency and the level of the constraints enforced by the applied policy. For example, if the policy aims at guaranteeing minimum data-rates, then the higher the summation of these rates over all users is as compared to system capacity, the more difficult fulfiling them becomes, leading to more resources being used, and consequently to a lower spectral efficiency.

The MT scheduler is not able to control user performance and this is clearly visible by observing the probability density function of the average user throughput (Figure 7) meaning that users' average throughputs are highly scattered.

Figure 8 depicts the system *satisfaction level* (defined in Equation (6)), provided by the MT scheduling algorithm as a function of the LT-US for different values of mRTR. Even if the satisfaction level is something strictly related to minimum QoS requirements and MT does not explicitly consider the mRTR in the resource allocation process, performance indicators are also computed in terms of user satisfaction, assuming that every subscriber could trace and evaluate the actual

DOI: 10.1002/wcm

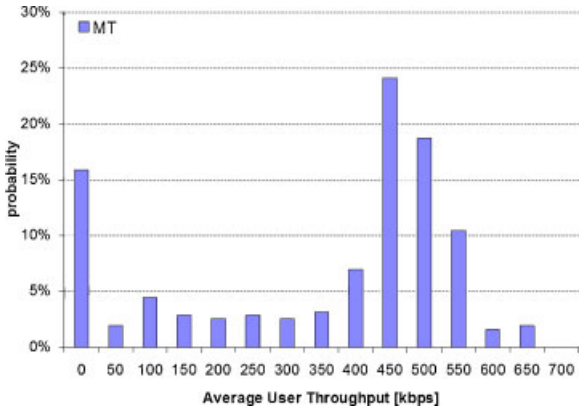


Fig. 7. Average user throughput probability density for maximum throughput scheduling.

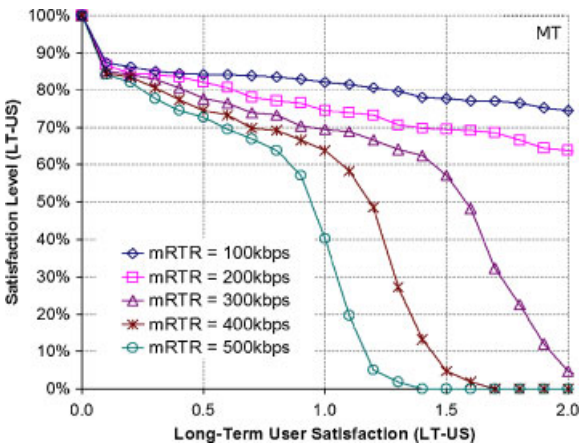


Fig. 8. System satisfaction level *versus* long-term user satisfaction using MT scheduling and for several values of the mRTR.

quality of the delivered service with respect to the subscribed one. This is essential in order to have a unified comparison over all scheduling approaches analysed.

Setting mRTR to 100 kbps the system results to be in a non-saturated state because the capacity necessary to fully support this level of QoS is 2.1 Mbps, which corresponds to 27% of the maximum achievable system throughput. Even if under these conditions the mRTR could be easily supplied for all users, about 19% of them measure an inadequate service level on the long term. Moreover, most of these users experience an LT-US close to 0, which confirms that MT systematically excludes some users from accessing to the radio channel. The percentage of

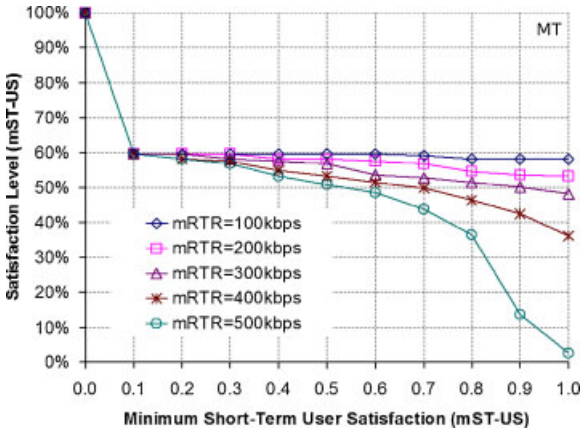


Fig. 9. System satisfaction level *versus* minimum short-term user satisfaction using MT scheduling and for several values of the mRTR.

unsatisfied users (LT-US < 1) increases up to 60% for mRTR $= 500$ kbps and this effect is even more evident when considering the mST-US, shown in Figure 9.

In a real system the chance that admitted users are not able to access the radio channel should be avoided at least for all non-BE subscribers. In terms of WiMAX specific traffic classes, the MT scheduler could only be used for the BE traffic class.

A commonly adopted concept used when addressing this issue is *fairness* [31,32]. *Fair* scheduling aims at finding a trade-off between throughput maximization and system equity, meaning that all users are allowed to access the radio channel; the fairness degree is, in a certain sense, a measure of the differences between their rates. It is important to note that fairness does not necessarily mean QoS: in fact, QoS deals with minimum guarantees while fair scheduling does not consider explicit constraints on minimum perceived performances.

A classic fair scheduling algorithm from literature is the PF rule, which provides users with a rate proportional to their average channel quality. Even if PF does not consider mRTR during the resource allocation process, a more fair behaviour can be observed by users from both the satisfaction level and the average throughput pdf (Figures 10,11 and 12).

However, the spectral efficiency is reduced due to the increased usage probability of robust MCSs. Figure 13 evidences that during the scheduling process the BS is forced to allocate resources using non-optimal MCSs.

The PF scheduling is obtained using a logarithmic utility function, having the first derivative

DOI: 10.1002/wcm

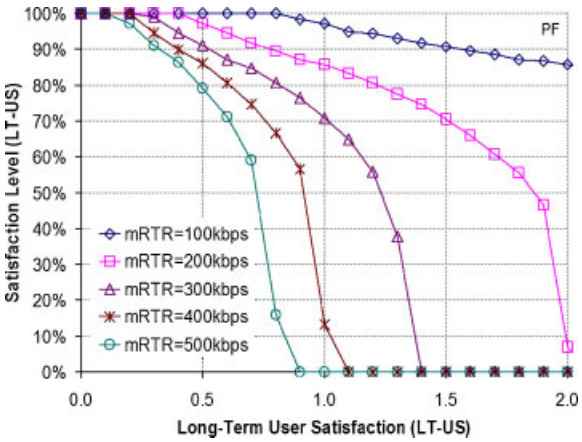


Fig. 10. System satisfaction level *versus* long-term user satisfaction using PF scheduling, for several values of mRTR.

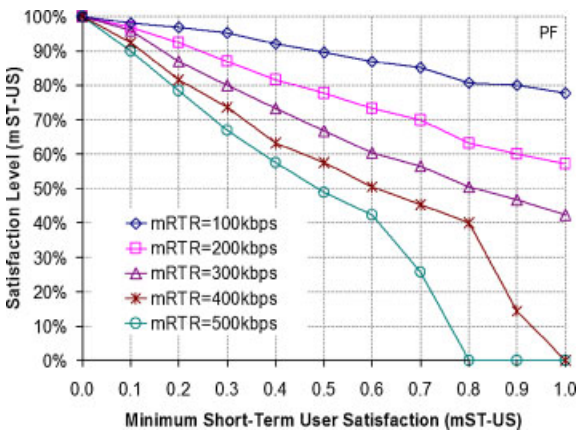


Fig. 11. System satisfaction level *versus* minimum short-term user satisfaction using PF scheduling, for several values of mRTR.

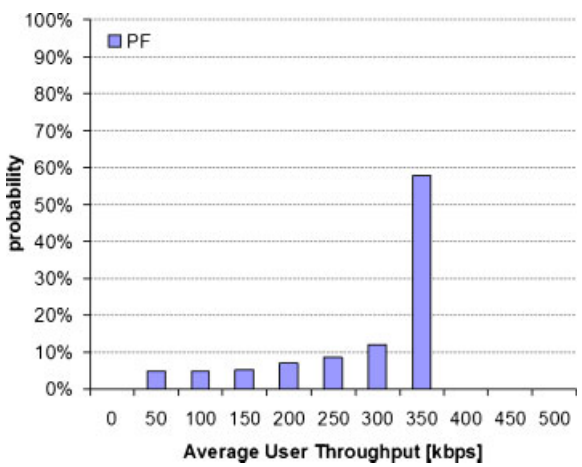


Fig. 12. Average user throughput probability density for proportional fair.

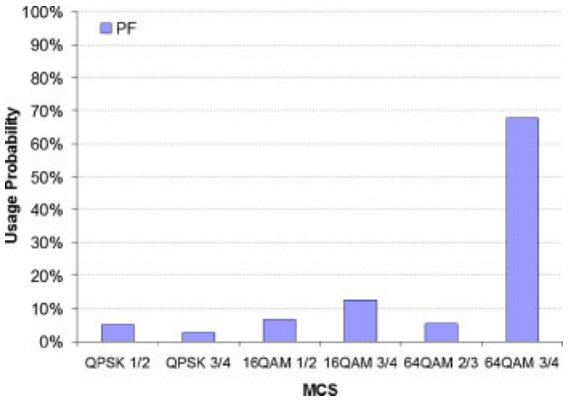


Fig. 13. PF MCS usage probability distribution.

(marginal utility):

$$U_i'(r_i) = \frac{1}{r_i} \tag{8}$$

where r_i is the average rate of the user i over the last time window TW = 0.5 s.

According to the utility-based scheduling approach, the marginal utility approximates the instantaneous profit (i.e. the dimension of the step taken in the direction of maximizing the objective function) that the scheduler could have for each additional bit transmitted by the user. In order to maximize the overall system utility, the scheduling metric used to assign each OFDMA slot in the frame has to be calculated by multiplying the current marginal utility with the instantaneous CQI of the user, i.e. the bit-per-slot associated with the currently supported MCS.

In case of a logarithmic utility function, the resulting metric is exactly the ratio between the instantaneous and the average data rate of the user; by using this function the scheduler will provide all users an average throughput proportional to their average channel quality. In particular, when the average rate of a user decreases, the marginal utility grows very quickly and the utility-related term becomes dominant in the scheduling metric. On one hand this provides a certain degree of fairness among users, in the long term, but on the other hand the BS is forced to schedule the user even though it experiences a lower channel quality, which requires the use of a non-optimal MCS in terms of transmission efficiency. This results in a measured average system throughput of 6.8 Mbps (spectral efficiency 1.22 bps/Hz), which means a loss of about 13% with respect to MT.

Stressing the concept of fairness, it is possible to determine the system throughput when perfect fairness

DOI: 10.1002/wcm

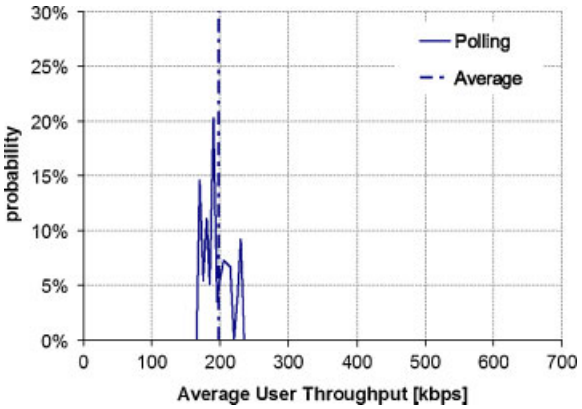


Fig. 14. Average user throughput probability density for Polling.

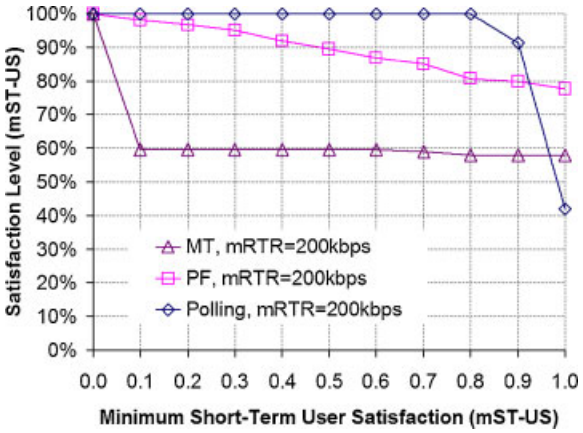


Fig. 15. System satisfaction level *versus* minimum short-term user satisfaction comparing MT, PF and Polling with mRTR = 200 kbps.

is enforced among users through a Polling scheduling algorithm, which periodically polls the queues of all active users and assigns to each one exactly the number of slots necessary for transmitting one packet. This way the available capacity is perfectly divided among all active users and the scheduler maximizes fairness, without considering any spectral efficiency optimization. Under these conditions the measured average system throughput is 4.4 Mbps (about 43% throughput loss), but all users are provided about 200 kbps (Figure 14).

Using this algorithm, the QoS constraints can be approximately fulfiled for all users for values of mRTR of up to 200 kbps. This value strongly depends on the particular simulation parameters (e.g. the number of simulated users); the scheduler can not control the rate of users in order to offer QoS guarantees on the mRTR. Considering all this, the Polling algorithms could be used as a benchmark for obtaining the minimum system performances when all users are provided the same rate, resulting in the highest possible loss in terms of spectral efficiency. Finally, maintaining the same simulation hypothesis for the other algorithms, mRTR = 200 kbps could represent the system saturation point.

In Figure 15 the satisfaction level curves as a function of the mST-US obtained for MT, PF and Polling are compared for mRTR = 200 kbps . From the plot it is clear that only Polling is able to provide minimum requirements by enforcing full fairness [§].

[§]In Figure 15 the maximum mST-US value for which all users are satisfied is 0.8 because mRTR = 200 kbps is obtained averaging values in different simulation runs that fluctuate depending on each particular simulated topology.

5.3. In the Search of QoS Guarantees

The way to guarantee QoS is to directly customize marginal utility functions so that the minimum QoS requirements are explicitly considered during the scheduling process. From the user's perspective, this approach allows differentiating the scheduler behaviour depending on what the current user satisfaction degree is while, from the overall system throughput viewpoint, the scheduler can allocate resources in order to satisfy minimum requirements and, only when minimum QoS is provided for all users, the spare capacity could be employed following some other criteria, e.g. throughput maximization. As an example consider the following marginal utility function (referred as *PF+MT*):

$$U_i'(r_i) = \begin{cases} A \cdot \frac{\text{mRTR}_i}{r_i} & \text{if } r_i \leq \text{mRTR}_i \\ 1 & \text{if } r_i > \text{mRTR}_i \end{cases}$$

When the average throughput r_i is below the mRTR the scheduler assumes that QoS is not satisfied for user i and calculates the scheduling metric in a PF way. However, when the minimum rate constraint is fulfiled, the utility function is independent of the average rate and users are ordered on the base of their CQI (like maximum throughput). The A parameter in the formula is essential in order to give strict priority to users below mRTR with respect to those already satisfied. In particular this parameter, which corresponds to the marginal utility value for an average rate exactly equal to the mRTR, should guarantee that the scheduling metrics of any satisfied users are lower than the minimum possible value for the unsatisfied ones. This

DOI: 10.1002/wcm

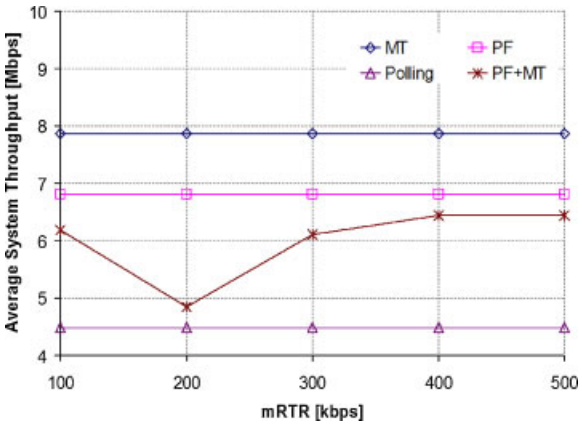


Fig. 16. System throughput *versus* mRTR for different algorithms.

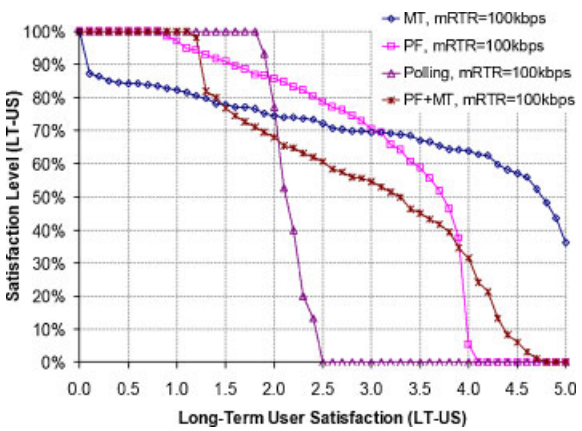


Fig. 17. System satisfaction level *versus* long-term user satisfaction with mRTR = 100 kbps.

condition holds if

$$A \geq \frac{\text{bps}_{64\text{QAM}3/4}}{\text{bps}_{\text{QPSK}1/2}}$$

Figure 16 shows the system throughput for various scheduling algorithms as a function of the mRTR. For low system load (e.g. mRTR = 100 kbp) the system throughput is higher than Polling and, at the same time, all users in the system are satisfied both on a long as well as on a short time frame, as shown by Figures 17 and 18.

From the user satisfaction plots we can deduce that, when the network load is below the saturation point, the PF+MT algorithm behaves as expected assuring that minimum requirements are satisfied for all users and using the spare capacity for system throughput maximization purposes. In fact all users have both minimum short-term and LT-US higher than 1 and

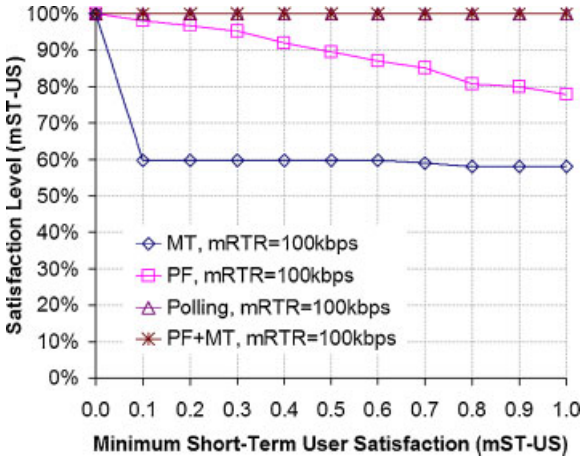


Fig. 18. System satisfaction level *versus* minimum short-term user satisfaction with mRTR = 100 kbps.

users with best channel quality reach an LT-US of up to 5.

The saturation point is reached for mRTR = 200 kbps; under these conditions the PF+MT algorithm performs better than both PF and MT, assuring that all users experience a good mST-US (up to about 0.6), while 80% of users reaching an mST-US of 1 (see Figure 19). At the same time the achieved system throughput is higher than Polling.

When mRTR further increases, the system goes in saturation and radio resources are not enough for providing QoS to all users. This results in PF+MT performing similarly to the pure PF from the point of view of both system throughput and user satisfaction. In fact, when there is a lack of resources, all users spend most time in the first zone of the utility function where the algorithm

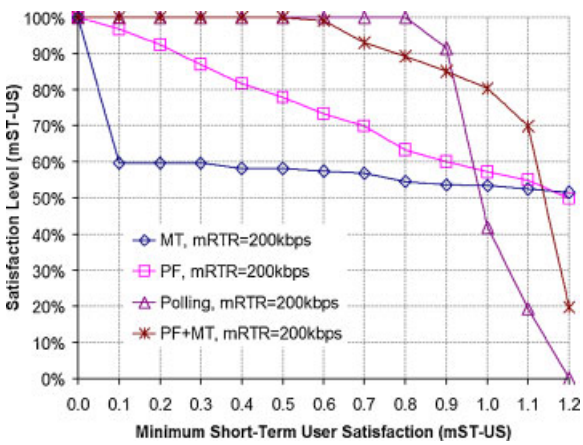


Fig. 19. System satisfaction level *versus* minimum short-term user satisfaction for mRTR = 200 kbps.

DOI: 10.1002/wcm

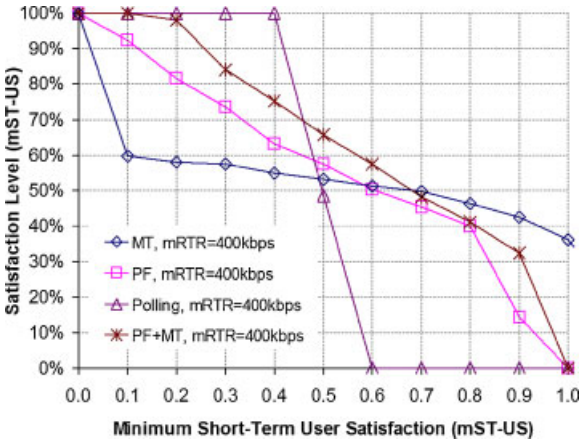


Fig. 20. System satisfaction level *versus* minimum short-term user satisfaction for mRTR = 400 kbps.

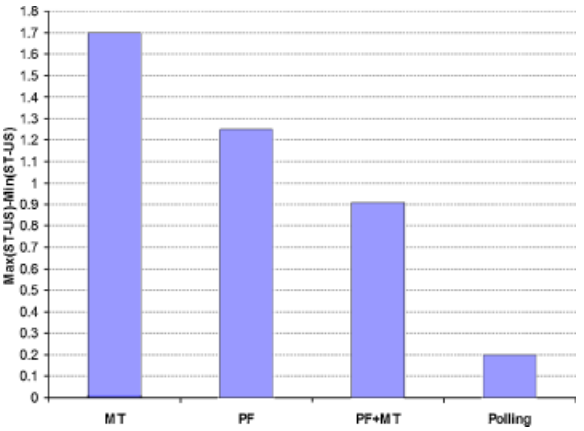


Fig. 21. Difference between overall maximum and minimum mST-US for mRTR = 300 kbps.

works as PF. However, the strict priority assigned to unsatisfied users gives slightly better performances. Figure 20 shows the similar behaviour of such algorithms.

In a real system, an overload condition can be caused by a call admission control (CAC) mechanism malfunction as well as traffic peaks and fluctuations. In that case, a robust scheduling policy should guarantee a *smooth performance degradation* in order to avoid the starvation of users with poor average channel conditions.

Moreover, a robust scheduling policy can allow the CAC to work very close to the real system capacity without any preventive security margin by relying on the scheduler for managing and absorbing instantaneous traffic peaks. When there is a lack of resources, all analysed algorithms, except Polling, intrinsically give priority to users with best channel condition because of the channel-dependent contribute (CQI) in the scheduling metric and only the design of fair utility functions can mitigate this problem. Furthermore, when network load exceeds the saturation point, even fair utility functions become substantially ineffective. Considering the mST-US, for example, and taking the minimum and the maximum between all simulated users, the difference between these two values is, in a certain way, inversely proportional to the *smoothness* of each scheduling algorithm. This difference is shown in Figure 21 for mRTR = 300 kbps.

In other words the higher the gradient of the mST-US curve is, the smoother the performance degradation provided by the scheduling algorithm is. Obviously, Polling offers the most homogeneous service degradation but, due to the fact that this algorithm does not consider any throughput-related criterion, system throughput loss exceeds 40%.

By exploiting the time dimension it is possible to improve the performance of the scheduler. Time diversity among users guarantees that peaks in channel quality occur at different and independent time instants. The introduction of a metric able to favour users when their channel quality experiences a peak with respect to the average could lead to two main advantages: firstly the mRTR is provided to users with poor channel quality, exploiting their most efficient transmission capabilities and thus minimizing the bandwidth necessary for minimum QoS requirements fulfilment, secondly this approach guarantees a smooth performance degradation because users with best absolute CQI are not privileged thanks to the relative peak-to-average (PTA) metric used.

In principle, this concept can be applied to every utility function. In this work the PF+MT scheduling approach is extended by multiplying each user's marginal utility with the ratio between its instantaneous CINR and a value averaged over a sliding time window, instead of the absolute CQI.

In Figure 22 the satisfaction level curves as a function of both long-term and minimum short-term user satisfaction are plotted for Polling, PF+MT and its extension with relative CQI (referred as PTA_PF+MT) simulated for mRTR = 300 kbps.

Satisfaction level for mST-US (solid lines) and LT-US (dotted lines) provided by Polling are very similar while the distance between the two curves of PTA_PF+MT indicates that the algorithm improves the system throughput at the detriment of the minimum experienced QoS level. In that case users will perceive a higher average satisfaction level but, unlike the Polling

DOI: 10.1002/wcm

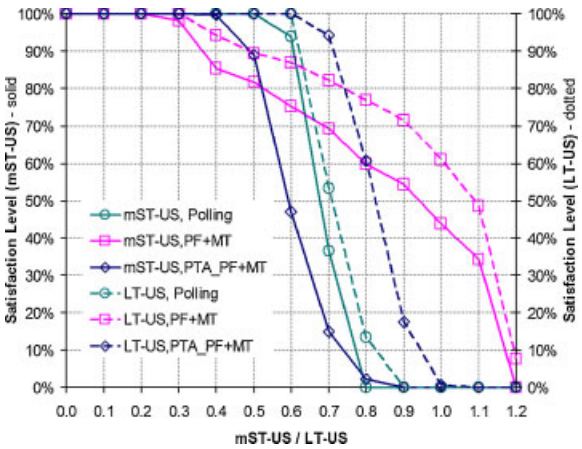


Fig. 22. System satisfaction level as a function of both LT-US and mST-US comparing Polling, PF+MT and its extension with peak-to-average relative metric (PTA_PF+MT) with mRTR = 300 kbps.

case, there will be a significant variability between the QoS levels provided in different time windows which causes the mST-US to shift to the left in the plot; but in any case the mST-US always remains above 0.4.

Moreover, as expected, using the peak-to-average relative metric when the system is saturated, the smoothness is improved and all users will experience nearly the same performance degradation. This consideration becomes clear by observing the difference between overall maximum and minimum mST-US measured by users, which is shown in Figure 23.

Figure 24 compares the time behaviour of PT+MT and PTA_PF+MT considering the same two users, u_1

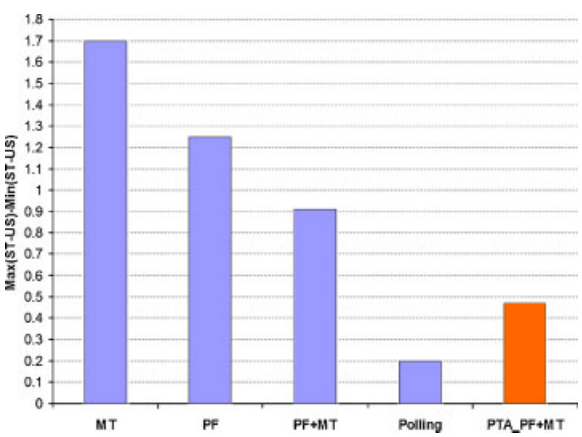


Fig. 23. Difference between overall maximum and minimum mST-US for mRTR = 300 kbps.

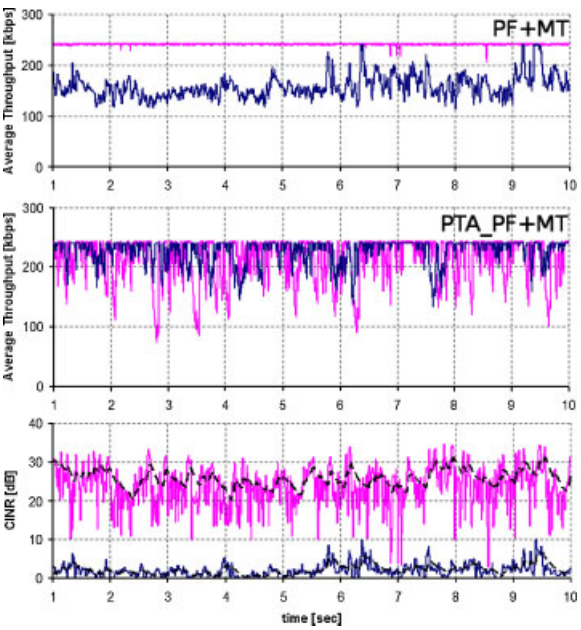


Fig. 24. Average throughput for PF+MT, PTA_PF+MT and channel quality, respectively, compared for two users with mRTR = 200 kbps.

and u_2, whose channel quality is depicted in the plot below. Figure 25, instead, shows the corresponding user satisfaction for each simulated time window. In this case mRTR is set to 200 kbps and the system is close to saturation. Using the absolute CQI metric, the PT+MT algorithm always gives priority to the user with the best channel conditions, while the second user

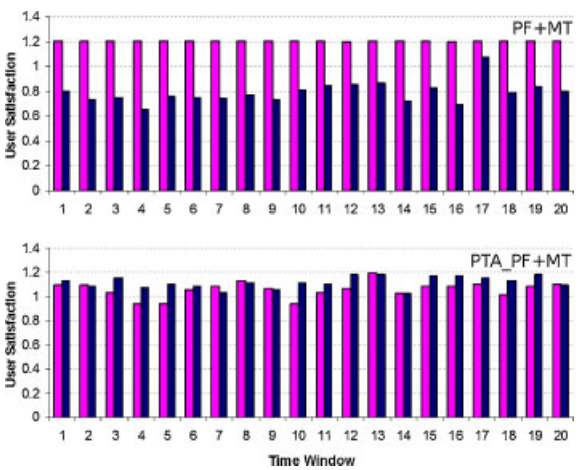


Fig. 25. US for each time window (0.5 s) for the two users with mRTR = 200 kbps.

DOI: 10.1002/wcm

Table VIII. User satisfaction comparing PF+MT and PTA_PF+MT.

User satisfaction	PF+MT	PTA_PF+MT
LT-US (u_1; u_2)	(1.2; 0.79)	(1.06; 1.12)
mST-US (u_1; u_2)	(1.2; 0.66)	(0.94; 1.03)

reaches the mRTR only in one time window[‖]. Using the peak-to-average metric instead, the PTA_PF+MT algorithm tends to satisfy all users in most time windows and both mST-US and LT-US, shown in Table VIII, are enhanced and close to 1.

The behaviour of the algorithms studied could be summarized as follows:

- **MT:** Provides the highest spectral efficiency. However, under high traffic demand, on average 46% of users do not reach their minimum data-rate guarantees on the long term; on the short term, 63% of them fail to reach the provisioned rates. Even under a very light system load, 19% of users measure an inadequate long-term service level.
- **PF:** Achieves a compromise between high spectral efficiency (13% lower than MT) and a certain degree of fairness. In a non-saturated system, the performance of the algorithm is good, leaving unsatisfied only 14% of users on the long term. When the traffic was high, the results were not that positive anymore, with 87% of users being left unsatisfied on the long term.
- **Polling:** Maximizes fairness at the expense of achieving poor spectral efficiency (43% drop as compared with MT). All users are provided a data rate of about 200 kbps.
- **PF+MT:** The benefits of this extension can be seen in the case of a non-saturated system, because it can better exploit the residual capacity left after providing the minimum data rate requirements to users. In this case all users are satisfied both on a long term and on a short term basis, and the system throughput is higher than Polling. When the system is saturated, the behaviour is similar to the one of the PF algorithm, maintaining fairness among users. A slight improvement can be noticed due to the use of strict priorities for unsatisfied users.
- **PTA_PF+MT:** By including in the utility function a term reflecting the relative channel quality with respect to its average value (over a sliding time window) for each user, a significant performance improvement can be noticed. Using this algorithm, all users experience a similar performance degradation in the case of a saturated system. When the system load is below capacity, this method achieves a satisfaction level close to 100% for all users.

[‖]The average rate of first user is close to 1.2 · mRTR due to the scheduler parametrization which requires to set the mRTR parameter slightly above the real minimum rate required. This is essentially due to the system overheads and time filter inaccuracy.

6. Conclusion

In this paper we have presented a detailed performance evaluation of utility-based scheduling schemes in 802.16e systems. Since radio resource management algorithms are not part of the standard, scheduling schemes are an important mechanism allowing vendors and providers to differentiate the service offer. The presented results are useful to help system designers select the most appropriate parameter settings to tailor the service to customer needs.

Figures 26 and 27 summarize the tradeoff of all the algorithms that we have presented, in both non-saturated (mRTR = 100 kbps) and saturated (mRTR = 300 kbps) network conditions. It can be clearly seen that, in order to meet user demands in terms of minimum data rate guarantees, a system operator needs to sacrifice the total system throughput. Depending on the stringency and on the level of these QoS requirements, the amount of resources that remains in each frame to be used by the operator for other purposes, such as maximizing system efficiency,

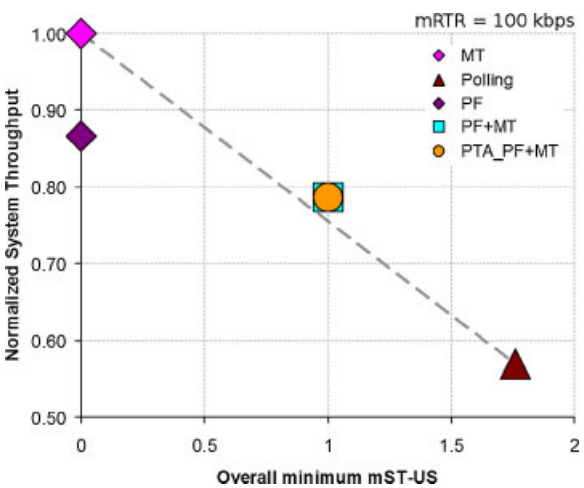


Fig. 26. Overall minimum mST-US *versus* average System Throughput for mRTR = 100 kbps.

DOI: 10.1002/wcm

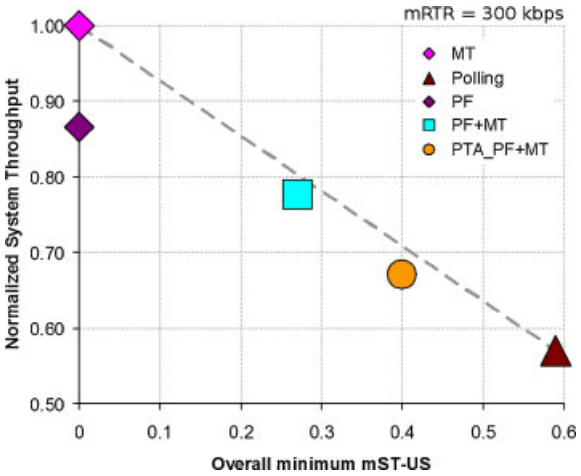


Fig. 27. Overall minimum mST-US *versus* average System Throughput for mRTR = 300 kbps.

could be extremely low, leading to reduced spectral efficiency. However, it should be noted that this loss in total throughput can be diminished by using efficient algorithms that exploit the properties of frequency, time and user diversity of the WiMAX system. By leveraging these properties, a scheduling algorithm could drastically increase the total system capacity, being able both to provision QoS and to increase the system efficiency at the same time.

References

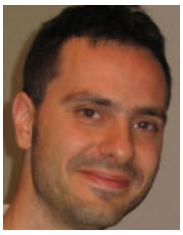
1. Fishburn P. *Utility Theory for Decision Making*. Wiley: New York, 1970.
2. Song G, Li Y. Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks. *IEEE Communications Magazine* 2005; **43**(12): 127–134.
3. IEEE Std 802.16-2004. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems.
4. IEEE Std 802.16e-2005. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands.
5. WiMAX Forum. Mobile WiMAX—Part I: A Technical Overview and Performance Evaluation, August 2006.
6. Ghosh A, Wolter DR, Andrews JG, Chen R. Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential. *IEEE Communications Magazine* 2005; **43**(2): 129–136.
7. Wang F, Ghosh A, Sankaran C, Fleming PJ, Hsieh F, Benes SJ. Mobile WiMAX systems: performance and evolution. *IEEE Communications Magazine* 2008; **46**(10): 41–49.
8. Yaghoobi H. Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN. *Intel Technology Journal* 2004; **8**: 201–214.
9. Riato N, Serrelli F, Sala A, Capone A. Interference Mitigation Strategies for WiMAX Networks. In *Proceedings of the IEEE Wireless Communication Systems*, Trondheim, Norway, October 2007; 175–179.
10. Riato N, Sorrentino S, Franco D, Masseroni C, Rastelli M, Trivisonno R. Impact of mobility on physical and MAC layer algorithms performance in Wimax system. In *Proceedings of the IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, Athens, Greece, September 2007; 1–6.
11. Pitic R, Capone A. An opportunistic scheduling scheme with minimum data-rate guarantees for OFDMA, In *Proceedings of the IEEE Wireless Communications and Networking Conference*, Las Vegas, NV, USA, March 2008; 1716–1721.
12. Mehrotra A. *GSM System Engineering*. Artech House, Inc.: Norwood, MA, USA, 1997.
13. Graja H, Perry P, Todinca D, Murphy J. Novel GPRS simulator for testing MAC protocols. In *Proceedings of the 3G Mobile Communication Technologies Conference*, 2003; 409–412.
14. Kim H, Kim K, Han Y, Yun S. A proportional fair scheduling for multicarrier transmission systems. In *Proceedings of the IEEE Vehicular Technology Conference*, 2004; 409–413.
15. Knopp R, Humblet PA. Information capacity and power control in single-cell multiuser communications. In *Proceedings of the IEEE International Conference on Communications*, Seattle, WA, USA, June 1995; 331–335.
16. Viswanath P, Tse DNC, Laroia R. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory* 2002; **48**(6): 1277–1294.
17. Andrews M, Kumaran K, Ramanan K, Stolyar A, Vijayakumar R, Whiting P. CDMA data QoS scheduling on the forward link with variable channel conditions. *Bell Labs Technical Memorandum* , April 2000.
18. Andrews M, Kumaran K, Ramanan K, Stolyar A, Whiting P, Vijayakumar R. Providing quality of service over a shared wireless link. *IEEE Communications Magazine* 2001; **39**(2): 150–154.
19. Shakkottai S, Stolyar A. Scheduling for multiple flows sharing a time-varying channel: the exponential rule. *American Mathematical Society Translations, Series 2,* (*A volume in memory of F. Karpelevich*) 2002; **207**: 185–202.
20. Badia L, Baiocchi A, Todini A, *et al.* On the impact of physical layer awareness on scheduling and resource allocation in broadband multicellular IEEE 802.16 systems. *IEEE Wireless Communications* 2007; **14**(1): 36–43.
21. Huang C, Juan HH, Lin MS, Chang CJ. Radio resource management of heterogeneous services in mobile WiMAX. *IEEE Wireless Communications* 2007; **14**(1): 20–26.
22. Niyato D, Hossain E. A queuing-theoretic and optimization-based model for radio resource management in IEEE 802.16 broadband wireless networks. *IEEE Transactions on Computers* 2006; **55**(11): 1473–1488.
23. Chen J, Wang CC, Tsai FCD, *et al.* The design and implementation of WiMAX module for ns-2 simulator. *ACM International Conference Proceeding Series*, 2006.
24. The Network Simulator—ns. Available online at: http://www.isi.edu/nsnam/ns/.
25. Baum DS, Salo J, Del Galdo G, Milojevic M, Kysti P, Hansen J. An interim channel model for beyond-3G systems. In *Proceedings of the IEEE VTC'05*, Stockholm, Sweden, May 2005.
26. Baum DS, Salo J, Milojevic M, Kysti P, Hansen J. MATLAB implementation of the interim channel model for beyond-3G systems (SCME), May 2005. Available online at: http://www.tkk.fi/Units/Radio/scm/.
27. Ben-Shimol Y, Kitroser I, Dinitz Y. Two-Dimensional Mapping for Wireless OFDMA Systems. *IEEE Transactions on Broadcasting* 2006; **52**(3): 388–396.
28. IEEE C802.16e-05/141r3. CINR measurements using the EESM method, April 2005. IEEE 802.16e contribution. Available online at: http://ieee802.org/16.

DOI: 10.1002/wcm

29. Caprara A, Monaci M. On the two-dimensional Knapsack Problem. *Operational Research Letters*, 2004.
30. Fayard D, Zissimopoulos V. An approximation algorithm for solving unconstrained two-dimensional knapsack problems. *European Journal of Operational Research* 1995; **84**(3): 618–632.
31. Wengerter C, Ohlhorst J, von Elbwart AGE. Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA. *IEEE 61st Vehicular Technology Conference*, June 2005; 1903–1907.
32. Lu S, Bharghavan V, Srikant R. Fair scheduling in wireless packet networks. *SIGCOMM*, September 1997.

Authors' Biographies

Razvan Pitic received his Dipl.-Ing. degree in Telecommunications in 2005 from Technical University of Cluj-Napoca, Romania. After a short experience as a network management engineer with Alvarion Romania, he joined the Ph.D. program at Politecnico di Milano, Italy, which he completed in Jan 2009. His doctoral dissertation was focused on packet scheduling algorithms for emerging OFDMA-based networks, with a special interest in frequency selective scheduling in the context of IEEE 802.16e networks. Currently he is a post-doc researcher at RFId Solution Center, Politecnico di Milano, where he is working on applications of sensor networks and RFID systems for identification/localization purposes, as well as on projects dealing with technology transfer from academia to industry.



Federico Serrelli received the M.S. degree in Telecommunications Engineering from the Politecnico di Milano University in 2005. From June 2005 to September 2008 he was research assistant at the Advanced Network Technologies Laboratory (ANTLab) of Politecnico di Milano where he worked on radio resource management algorithms for WiMAX. Between 2006 and 2007 he collaborated with Nokia Siemens Networks on the IEEE 802.16d/e performance evaluation and product development support. He is currently working as Software Engineer and information and communication technology (ICT) consultant giving support to private companies in the development of telecommunication and IT-related applications and services. His research interests focus on wireless access networks, advanced RRM techniques, *ad hoc* networks and routing, IP networking and service development, distributed computing and architectures.

Simone Redana received his M.Sc. in Communication Engineering in 2001 and his Ph.D. in 2005 from Politecnico di Milano. From June 2005 to May 2006 he worked in Azcom Technology as consultant for Siemens Communication. In June 2006 he jointed Siemens Communication in Milan Italy, and from April 2007 he was with Nokia Siemens Networks Italy. Since January 2008 he is with Nokia Siemens Networks in Munich Germany. From June 2006 to December 2008 he contributed to the EU WINNER II Project and to the Eureka Celtic project WINNER+. From July 2008 he is actively working on LTE-Advanced.

Antonio Capone is an Associate Professor at the Information and Communication Technology Department (Dipartimento di Elettronica e Informazione) of the Technical University of Milan (Politecnico di Milano), where he is the director of the Advanced Network Technologies Laboratory (ANTLab). His expertise is on networking and his main research activities include protocol design (MAC and routing) and performance evaluation of wireless access and multi-hop networks, traffic management and QoS issues in IP networks, and network planning and optimization. On these topics he has published more than one hundred peer-reviewed papers in international journal and conference proceedings, and holds five patents.

He received the M.S. and Ph.D. degrees in electrical engineering from the Politecnico di Milano in 1994 and 1998, respectively. In 2000 he was visiting professor at UCLA, Computer Science department. He currently serves as editor of Wireless Communications and Mobile Computing (Wiley), Computer Networks (Elsevier), and Computer Communications (Elsevier). He was guest editor of a few journal special issues and served in the technical program committee of major international conferences (including Mobicom, INFOCOM, SECON, MASS, Globecom, ICC, LCN, Networking, WoWMoM), as Technical Program Chair of Ifip MEDHOCNET 2006, Poster&Demo co-chair of SECON 2009, and publicity chair of MOBIQUITOUS 2007.

He is currently involved in the scientific and technical activities of several national and European research projects, and he leads several industrial projects. He is a Senior Member of the IEEE.

DOI: 10.1002/wcm